Final Exam Required Section

For this exam, you will need to submit your responses in a form on Blackboard. In addition, there will be a blank at the end for you to fill in the name of students you collaborated with. You will also need to submit your code via a separate link. Answers that are not supported by code will not receive full credit. **You should report all answers to 4 decimal places unless otherwise specified.**

This is a take-home final exam. You are permitted to work with up to 2 other students on this portion of the exam. This means a closed group of up to 3 people. Not one person who works with 2 other people, and then those 2 other people work with 2 other people. This will be explained in class and claiming not to understand the policy is not an acceptable reason for violating it. Be aware that unauthorized collaboration on a take-home exam or paper is a Level III Honor Code violation with one of the following consequences:

Deferred Suspension, Suspension, Indefinite Suspension, or Permanent Dismissal

It is much easier to recover from a low grade on an assignment or in a course than it is to recover from being expelled from the university.

Part 1: Earthquake Data

The data for this problem was sourced from https://github.com/fivethirtyeight/data/tree/master/, and was used in the article https://fivethirtyeight.com/features/the-rock-isnt-alone-lots-of-people-are-worried-about-the-big-one/. We will assume that the individuals from who data was collected were randomly sampled. You have been given an adaptation of the original data in the file earth_orig_dropped.csv – you should NOT download the original data, it is different.

Q1: Using this data as is (that is, without transforming it), what distance/dissimilarity metric(s) could we have calculated with Python to look at relationships between individuals based on the variables Worried (how worried are you about earthquakes), Worried_Big_One (how worried are you about the Big One), Big_One_Lifetime (do you think the Big One will occur in your lifetime), and Preparations (have you taken any precautions for earthquakes?).

Q2: How would you re-code this data, specifically looking at the 4 columns indicated in Q1? Describe in 2-4 sentences how you would re-code the variables. “How” does not mean the code you would use, it means what the variables would look like after you re-code them, and your schema/rationale for assigning different values.

Q3: Using this re-formatted data, which distance metric(s) could you now choose to explore differences between individuals in addition to what you chose for Q1 (that is, don’t repeat your Q1 answers)?

Q4: Justify your choice(s) for Q3 and why it/they might be preferable to the metric(s) you cited in Q1 in 1-3 sentences.

Q5: You perform PERMANOVA using a version of the data in earth_orig_dropped and obtain the following results:

```
Call:
adonis(formula = earth ~ earth_metadata$Experience, permutations = 999, method='euclidean')

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

                          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
earth_metadata$Experience  2     320.8 160.413  40.873 0.07736  0.001 ***
Residuals                975    3826.5   3.925         0.92264
Total                    977    4147.3                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which of the following statements apply?

a) The null hypothesis for the PERMANOVA is that the mean amount of worry is the same.
b) At a 1% level of significance, we conclude that having experienced an Earthquake causes people to worry more about earthquakes.
c) There is nearly 1000 times more variability between individuals who have varying levels of experience with earthquakes than among individuals with the same level of experience.
d) There are less than 5 permutations of the data that provides a Pseudo-F statistic which is less than what we saw from the actual data.
e) We should have used Jaccard distance since the data was categorical.
f) Over 90% of the variation in worry and preparation for earthquakes can be explained by an individual's prior experience with earthquakes.
g) Experience explains most of the differences we see in levels of worry and preparation.
h) There are two degrees of freedom for Experience because there are two experience levels.
i) This analysis must have been performed on re-coded data.
j) The null hypothesis is no difference in mean number of earthquakes experienced for individuals over varying levels of worry.
k) At a 5% level of significance we would conclude that worry/preparation for earthquakes is associated with having experienced an earthquake.
l) In the original data, individuals with the same degree of earthquake experience vary approximately 40 times less from each other than individuals with different degrees of earthquake experience.
m) We need to use a permutation test because of the small sample size.

Q6: Why was PERMANOVA used above as opposed to regular ANOVA? Explain in 1-3 sentences.

Q7: You suspect that region may be a confounder in the relationship between whether someone has experienced an earthquake and their level of worry/preparation (the variables used to make the distance matrix). How would you use the data you have to evaluate this? YOU DO NOT NEED TO DO THE ANALYSIS but describe how you would do the analysis to identify/control for confounding by region in 2-4 sentences. Explain how does not mean giving code, it means describe the methodology AND rationale behind it.

The article states that: It turns out the more you know about the San Andreas fault or the Yellowstone supervolcano — and if you don't know about the last one, you are in for a major treat — the more likely you are to be worried about "the Big One."

Q8: Make a figure to show the relationship between knowing about the Yellowstone Volcano and being worried about the Big One. Does your figure support the article's claim? Explain in 1-3 sentences.(You don't need to calculate statistics, answer this based on your figure). Upload your figure in Q9, and be sure that it shows the relationship the article discusses and is presented in a way that is easy to read and makes sense.

Q10: Conduct a statistical test to determine whether the level of knowledge of the Yellowstone supervolcano is associated with the level of worry about the Big One at a 1% level of significance. State the null hypothesis, name the test you used, justify your use of this test, and state a p-value. Based on your results, what do you conclude in the context of the problem?

The article also states that "People who think "the Big One" will happen in their lifetimes are about three times as likely to have taken precautions for a disaster, such as buying a survival kit or making an evacuation plan, according to the survey."

Q11: Make a table to display the relationship (in terms of count of individuals) between preparing for an earthquake ('Preparation') and thinking the Big One will happen in your lifetime ('Big_One_Lifetime'). Upload an image (screenshot is fine) of your table for Q12. Does your data table support the article's statement? Explain why or why not. Would you re-write their claim in any way for accuracy or specificity? Answer these questions in 2-4 sentences.

Q13: It said in the instructions to assume that this was a random sample. Read the article and evaluate the veracity of that statement in 1-3 sentences. (I.e. was it really a random sample? How do you know? What would be a better way to describe it?)

Part 2: Drug Use Patterns

This part of the assignment uses simulated data based on data available from the National Survey on Drug Use and Health. Use and dissemination of data from the survey is restricted, which is why the data for this assignment is simulated.

Suppose that the data in file drugs.txt was collected via a survey of randomly selected college students. Students filled out an anonymous questionnaire to indicate their usage (YES/NO) of particular substances.

Q14: Calculate a 99% confidence interval for the measure of association for the relationship between Opioid use and Alcohol use in this group of students. Interpret the point estimate in words. Use this interval to tell the null hypothesis of no association. Based on your interval, what do you conclude about Opioid use among those who use alcohol?

Q15: Is there an association between alcohol use and marijuana use in these students at the 5% level of significance? That is, are those who use alcohol more likely to also use marijuana? Provide an estimate for the appropriate measure of association and also a p-value for the hypothesis test you conducted. Justify the use of the specific test your chose. What do you conclude in the context of the problem?

Q16: Suppose that a news article uses this data and publishes the following result: "College students who use marijuana are at double the risk for opioid use as students who do not use marijuana." Do you agree with this result? Justify your answer in 2-3 sentences.

Part 3: Mixed bag

Q17: You conduct a survey to explore the relationship between international airline travel and acquisition of the flu. Of 200 randomly sampled individuals, 45 of them reported a positive diagnosis of the flu using a rapid influenza diagnostic test. Of those with the flu, 20 took an international flight in the six months prior. Of those without flu, 18 reported taking an international flight. Which of the following statements apply?

a. This study used blocking to decrease bias.
b. We are 95% confident that the true odds ratio for getting the flu for those who flew internationally versus those did not is between 1.04 and 2.57.
c. Since the flu is rare, we can estimate the risk ratio for flu as approximately 6 for those who fly internationally versus those who do not.
d. The results indicate that not flying internationally may be protective.
e. The 95% confidence interval for the odds ratio is borderline significant.
f. On average, odds ratio estimates will be within 0.39 of the true value of the odds ratio for a sample size of 200.
g. Confidence intervals for the OR can be calculated using Z since the sampling distribution for the odds ratio is normally distributed.
h. From our sample, we conclude at the 1% level of significance that the odds of getting the flu for those who flu internationally is significantly higher than for those who do not fly internationally.

Q18: Later on, some of the individuals who reported not having the flu ended up displaying symptoms. You find out (magically!) that of the individuals who tested negative for the flu, 80% were true negatives (i.e. they really didn't have the disease). This means that:

a) The prevalence of flu is high.
b) The specificity of the diagnostic test is 20%.
c) Misclassification bias was likely an issue in this study.
d) The test data is not normally distributed.
e) The negative predictive value is 80%.
f) If we limited the study to just these individuals, we could assess risk.
g) The probability of testing negative given that you have the flu is 20%.

Q19: You collect data on the number of shares of stock individuals own for the 30 companies that contribute to the Dow Jones Industrial Average to explore how various variables such as age, income, and occupation may affect the composition of stock portfolios. You obtain data on stock holdings in the following format (this is not all of the data, just a subset – consider that the metadata is stored elsewhere):

	ID_1	ID_2	ID_3	ID_4	ID_5
MMM 3M	28	6	50	5	100
AXP American Express	48	8	50	4	101
AAPL Apple	71	1	53	5	101
BA Boeing	23	5	54	5	100
CAT Caterpillar	49	5	54	1	100
CVX Chevron	91	9	59	1	101
CSCO Cisco	15	8	50	5	101
KO Coca-Cola	24	3	54	1	101
DIS Disney	25	6	60	5	0
DWDP DowDuPont Inc	68	5	55	2	100
XOM Exxon Mobil	45	7	50	5	101
GS Goldman Sachs	19	3	51	2	100
HD Home Depot	20	3	58	0	101

IBM IBM	35	5	56	3	101
INTC Intel	29	6	53	4	0
JNJ Johnson & Johnson	61	4	52	1	100
JPM JPMorgan Chase	53	0	56	3	101
MCD McDonald's	100	0	50	5	100
MRK Merck	30	1	60	1	101
MSFT Microsoft	11	3	54	2	101
NKE Nike	46	3	55	1	100
PFE Pfizer	42	10	51	1	101
PG Procter & Gamble	68	2	59	0	101
TRV Travelers Companies	50	2	54	5	101
UTX United Technologies	89	10	60	3	0
UNH UnitedHealth	29	6	55	4	101
VZ Verizon	47	9	50	1	101
V Visa	84	9	54	0	100
WMT Wal-Mart	0	5	57	0	0
WBA Walgreen	94	0	55	3	101

Each row is the name of the one of the stocks and each column corresponds to an individual in the study. The data in the table is the number of shares of stock each individual owns.
You wish to use this data to construct a distance/dissimilarity matrix for multi-dimensional scaling. Suppose the data above is stored in a matrix called `stocks`, and you will use pdist to construct the matrix as follows:

```
distance_matrix = scipy.spatial.distance.pdist(stocks.as_matrix())
```

Describe the steps you will need to take before running this command in order to ensure that you get a proper distance matrix in order to answer the original question. (You are not concerned with the metadata here, assume that is in the proper format for the question.) YOU DO NOT NEED TO WRITE CODE. Explain your answer in 3-5 sentences.

Q20: With respect to the example in Q19, which of the distance/dissimilarity metrics we discussed would be appropriate for this data?